



UNIVERSITY OF
ILLINOIS LIBRARY
AT URBANA-CHAMPAIGN
BOOKSTACKS

THE HECKMAN BINDERY, INC.
North Manchester, Indiana

KRL

BINDING COPY

H or V	JUST	FONT	SLOT	TITLE
H	CC	1W	22	BEBR
			21	FACULTY
			20	WORKING
			18	PAPER
H	CC	1W	8	1990
			7	NO. 1705-1718
H	CC	1W		330
				B385<"CV">
				no. 1705-1718
				cop. 2
H	CC	7W		<IMPRINT>
				U. of ILL.
				LIBRARY
				URBANA

PERIODICAL <input type="checkbox"/>	CUSTOM <input type="checkbox"/>	STANDARD <input type="checkbox"/>	ECONOMY <input type="checkbox"/>	THESIS <input type="checkbox"/>	NO. VOLS THIS TITLE	LEAD AT
BOOK <input type="checkbox"/>	CUSTOM <input type="checkbox"/>	MUSIC <input type="checkbox"/>	ECONOMY <input type="checkbox"/>	AUTH. 1ST <input type="checkbox"/>		
ACCOUNT	LIBRARY	NEW	RUB OR SAMPLE	TITLE I.D.	FOIL	COLOR
56672	001				6632	WRT 488
ACCOUNT NAME						
UNIV OF ILLINOIS						
ACCOUNT INTERNAL I.D.					ISSN.	
B01912400						
I.D. #2	NOTES	BINDING FREQUENCY	WHEEL	SYS. I.D.		
STX3			1	3		3920
COLLATING						
35						
ADDITIONAL INSTRUCTIONS						
Dept-STX3 Lot=#20 Item=151 HNM-1294						
1CR2ST3CR MARK BY # B4 911						
SEP SHEETS	PTS. BD. PAPER	TAPE STUBS	CLOTH EXT.	GUM	FILLER	STUB
POCKETS			SPECIAL PREP			LEAF ATT
PAPER	BUCK	CLOTH				
INSERT MAT	ACCOUNT LOT NO.			JOB NO.		
	#20			HV363		
PRODUCT TYPE	ACCOUNT PIECE NO.			PIECE NO.		
	11			151		
HEIGHT	GROUP CARD	VOL THIS TITLE				
11.2	1	2				
COVER SIZE						
V X 9						
						00124752

330
B385
No. 1717 COPY 2

STX

Convergence Analysis of Local Feature Extraction Algorithms

K. Hornik


*Technische Universität Wien
Vienna, Austria*

C.-M. Kuan

*Department of Economics
University of Illinois*



Bureau of Economic and Business Research
College of Commerce and Business Administration
University of Illinois at Urbana-Champaign



Digitized by the Internet Archive
in 2011 with funding from
University of Illinois Urbana-Champaign

<http://www.archive.org/details/convergenceanaly1717horn>

BEBR

FACULTY WORKING PAPER NO. 90-1717

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

December 1990

Convergence Analysis of Local
Feature Extraction Algorithms

K. Hornik

Institut für Statistik und Wahrscheinlichkeitstheorie
Technische Universität Wien, Vienna, Austria

and

C.-M. Hornik

Department of Economics
University of Illinois at Urbana-Champaign

The second author is grateful for support by the Investors in Business Education and the Bureau of Economic and Business Research of the University of Illinois.

Abstract

We investigate the asymptotic behavior of a general class of on-line Principal Component Analysis (PCA) learning algorithms, focussing our attention on the analysis of two algorithms which have recently been proposed and are based on strictly local learning rules. We rigorously establish that the behavior of the algorithms is intimately related to an ordinary differential equation (ODE) which is obtained by suitably averaging over the training patterns, and study the equilibria of these ODEs and their local stability properties. Our results imply in particular that local PCA algorithms should always incorporate hierarchical rather than more competitive, symmetric decorrelation, for reasons of superior performance of the algorithms.

1 Introduction

The ability to extract the main features inherent in complex, high-dimensional input data streams is of fundamental importance to many information processing systems. Such “dimensionality reduction” occurs e.g. as a preprocessing stage for efficient pattern recognition and classification, helps eliminating disturbing noise or information redundancy, and is necessary to allow for further transmission of the relevant information the input signal contains if not enough transmission channels are available.

Generally speaking, optimal feature extraction can be described as constructing a function F which compresses a d -dimensional input vector x into a p -dimensional output vector $y = F(x)$, where $p < d$ and usually $p \ll d$, such that, relative to some performance criterion, y contains as much information about x as possible. If the mean squared error of the best linear estimate of x given y (the “linear reconstruction error”) is used as a criterion, this leads to a statistical method called Principal Component Analysis (PCA); see Bourlard & Kamp (1988), Linsker (1988), Sanger (1989), Baldi & Hornik (1991). PCA is one of the simplest and most general purpose feature extraction techniques which extracts information by finding the directions in which the inputs exhibit most significant variation. The PCA outputs are given as $y = Ax$, where A is a $p \times d$ matrix such that the rows of A span the same subspace of \mathbb{R}^d as the eigenvectors associated with the p largest eigenvalues of the input covariance matrix, see e.g. Baldi & Hornik (1991).

A variety of PCA learning algorithms have been proposed within the last decade. If upon presentation of a new learning pattern x , we modify A according to

$$\Delta A = \gamma(yx' - yy'A), \quad y = Ax, \quad (1)$$

(in what follows, γ is the learning rate and $'$ denotes transpose), we obtain an algorithm introduced independently by Williams (1985) as the Symmetric Error Correction Algorithm (SEC), by Baldi (1988) as a symmetric simplification of the Back Propagation algorithm for a linear d - p - d architecture in autoassociative mode, and by Oja (1989) as the subspace algorithm. The Generalized Hebbian Algorithm (GHA) introduced by Sanger (1989) updates A by

$$\Delta A = \gamma(yx' - \text{lower}(yy'A)), \quad (2)$$

where the “lower” operator sets all entries above the main diagonal to zero.

Clearly, in both algorithms, the first additive term is just hebbian learning performing gradient descent on an energy function which maximizes the sum

of the output unit variances, whereas the second term tends to keep the weight matrix A constrained suitably; for the symmetric algorithm, it keeps AA' close to the p -dimensional identity matrix (Baldi & Hornik, 1991), and in Sanger's algorithm, it performs Gram-Schmidt orthonormalization on the rows of A , thereby hierarchically decorrelating the outputs. Both algorithms are, although only implicitly, already contained in Oja & Karhunen (1985). For the one-unit case ($p = 1$), they both reduce to the algorithm first introduced by Oja (1982) as a small learning rate first order approximation to hebbian learning with additional euclidean normalization of the weight vector; in the sequel, we shall always refer to this algorithm as "Oja's one-unit algorithm".

Recently, Földiák (1989) and Rubner & Tavian (1990) have introduced two new algorithms which are based on a combination of Oja's one-unit algorithm applied to each of the rows of A (or, roughly equivalent to that, hebbian learning with rowwise euclidean normalization) and some lateral inhibition mechanism designed for decorrelating the outputs. Hence, the network architectures employed for these algorithms utilize an additional set of connection weights accounting for decorrelation. Learning of A is strictly *local* in the sense that the modification of the i -th row of A only depends on the input x and the i -th network output; for the remaining weights, anti-hebbian learning is used, which again is a very simple, local rule. Due to the locality of these learning mechanisms, it has been argued that these algorithms be "biologically more plausible" than the subspace algorithm or GHA.

Our concern in the present paper is a rigorous analysis of the properties of these local PCA feature extraction algorithms for the case where a large number of training samples is available. Such an analysis is usually based on the claim that, assuming that the weight changes are sufficiently small, the sequence of weights generated by the learning algorithm can be approximated by the solution paths of an ordinary differential equation (ODE) which is obtained by "averaging over all patterns", and that the weights converge to the asymptotically stable equilibria of this ODE. We shall provide a precise result supporting this claim for the case where the training patterns are independent centered random variables with the same covariance matrix and the learning rates tend to zero at a suitable rate. In particular, we describe the appropriate averaging procedure for local algorithms based on a feedback architecture, such as Földiák's. In addition to that, we present a stability analysis of the equilibria of the ODEs associated with the algorithms. In particular, we shall establish that for symmetric lateral inhibition mechanisms between the outputs, as the architecture originally introduced in Földiák (1989), the desired limit points are *not*

asymptotically stable equilibria of the associated ODE. Therefore, hierarchical decorrelation, although clearly disallowing for “competition” and lacking symmetry, should be favored over symmetric decorrelation, for reasons of superior performance.

This paper is organized as follows. Section 2 introduces a general class of local PCA feature extraction algorithms which contains the ones introduced in Földiák (1989) and Rubner & Tavian (1990). Section 3 describes a precise method of associating an ODE to on-line learning algorithms. Results are given in section 4. Section 5 contains some additional remarks. All proofs are deferred to the appendix.

2 Local PCA Algorithms

One class of local PCA feature extraction algorithms, including the algorithm introduced by Rubner & Tavian (1990), can be described as follows, cf. Baldi & Hornik (1991), Kuan & Hornik (1990). Using an additional linear layer for decorrelation, the network output is computed as

$$y = QAx,$$

where Q is a $p \times p$ matrix which performs the decorrelation.

Upon presentation of a new learning pattern, A is updated using either hebbian learning with rowwise normalization, or, basically equivalent thereto if the learning rates are small, using Oja’s one-unit algorithm applied to each of the rows of A , which can compactly be written as

$$\Delta A = \gamma(yx' - \text{diag}(yy')A). \quad (3)$$

Here, the “diag” operator sets all off-diagonal entries to zero. In the sequel, we shall also use the “subdiag” operator, which sets all entries on or above the main diagonal to zero, and the “offdiag” operator, which sets all diagonal entries to zero.

Using the decorrelation mechanism proposed by Barlow & Földiák (1989), Q is written as $Q = I + W$ where I is the identity matrix and W is symmetric with zero diagonal and updated using the simple, anti-hebbian learning rule (as in the novelty filter of Kohonen (1984))

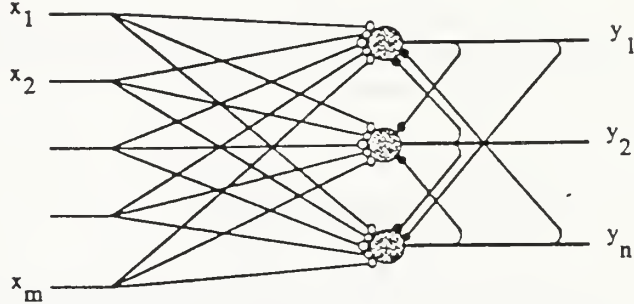
$$\Delta W = -\mu \text{offdiag}(yy'). \quad (4)$$

Alternatively, hierarchical (i.e. in some sense Gram-Schmidt type) decorrelation can be employed, which is accomplished upon writing $Q = I + W$, where now

W is subdiagonal (i.e., all entries of W which are on or above the main diagonal are zero), and updating W according to

$$\Delta W = -\mu \text{subdiag}(yy'). \quad (5)$$

The architecture introduced by Földiák (1989) is depicted below; white circles indicate hebbian, black circles anti-hebbian connections.



The network architecture proposed by Földiák

Here, the network outputs are the sum of weighted network inputs and the weighted feedback received from the other output units, such that, upon presentation of an input x , the outputs are updated according to

$$Ax + Wy \mapsto y, \quad (6)$$

where W is the $p \times p$ matrix of lateral connection strengths.

Initially, $W = O$ and A is “random”. Földiák suggests updating A and W according to rules (3) and (4), respectively, such that W is kept *symmetric* with zero diagonal throughout the learning process, and claims (p. 402) that when an input is presented to the network, the units settle to a stable state for which

$$y = Ax + Wy,$$

or

$$y = (I - W)^{-1}Ax.$$

However, this is not necessarily true (see also Baldi & Hornik, 1991). Let us write $y(k)$ for the network output after k updating cycles using (6), with fixed input x and initial output $y(0)$. Clearly,

$$\begin{aligned} y(k) &= Wy(k-1) + Ax \\ &\vdots \\ &= W^k y(0) + (I + W + \dots + W^{k-1})Ax. \end{aligned}$$

To ensure convergence of $y(k)$ to $(I - W)^{-1}Ax$ as $k \rightarrow \infty$, we thus need that all eigenvalues of W are less than one in absolute value, which is not guaranteed by the algorithm. Even if the algorithm is modified accordingly, it would still require infinitely many cycles to converge to the stable state, which is of course computationally infeasible for real-time applications. (Of course, the linear system $(I - W)y = Ax$ could be solved explicitly in finite time; but then the architecture is no longer self-contained, and the particularly attractive feature of performing only simple local computations is lost.)

Both problems can be overcome by using asymmetric (=hierarchical) decorrelation, i.e. using learning rule (5) rather than (4), which together with the initialization $W = O$ keeps W strictly subdiagonal throughout the learning process. In fact, in this case, $\lambda = 0$ is the only eigenvalue of W , and it is easily seen that the (i, j) -th entry of W^k vanishes if $i \leq j + k$ such that in particular, $W^k = O$ for all $k \geq p$ and

$$(I - W)^{-1} = I + W + \dots + W^{p-1}.$$

Hence, after p updating cycles, the stable state is reached, irrespective of the initial network output $y(0)$. Interestingly enough, it will be shown in section 4 that if asymmetric decorrelation is employed, then, during the learning period, it is enough to perform at least 2 cycles before updating the weight matrices, thereby making the algorithm “quicker”.

All algorithms introduced thus far are of the following general form. Upon presentation of a new learning pattern x , the network output y is updated according to

$$P(W)y + Q(W)Ax \mapsto y, \tag{7}$$

where $P(W)$ and $Q(W)$ are polynomials in the $p \times p$ matrix W which satisfy $P(O) = O$ and $Q(O) = I$, and then A and W are updated as

$$\Delta A = \gamma (yx' - \Phi(yy')A), \tag{8}$$

$$\Delta W = \mu \Omega(yy'), \tag{9}$$

where Φ and Ω are suitable linear (selection) operators on the space of all $p \times p$ matrices; as initializations, we take $W = O$, $y = 0$, and A as “random”. In fact, taking $\Omega \equiv O$ and Φ as the identity mapping, we obtain the subspace algorithm (1); the choice $\Omega \equiv O$ and $\Phi = \text{lower}$ gives Sanger’s GIIA (2). If both P and Ω are nonzero, the network outputs receive feedback from previous outputs.

All local algorithms use $\Phi = \text{diag}$. For the algorithm of Rubner & Tavian, $P \equiv O$, $Q(W) = I + W$, $\Phi = \text{diag}$ and $\Omega = -\text{subdiag}$. Real-time implementations based on Földiák’s architecture, which perform only a finite number, say k , of output updating cycles before updating the weight matrices, use $P(W) = W^k$, $Q(W) = I + W + \dots + W^{k-1}$, $\Phi = \text{diag}$, and $\Omega = -\text{subdiag}$ or $-\text{offdiag}$.

Of course, the above class of algorithms could be enlarged by allowing for more general functions P and Q ; for example, using Földiák’s original idea to let the outputs settle into the stable state before updating the network weights corresponds to the choice $P \equiv O$ and $Q(W) = (I - W)^{-1}$ (formally, the feedback is eliminated by stabilization). However, as already pointed out, this choice is computationally infeasible if combined with (4), and contained in the polynomial setting if W is kept subdiagonal using (5). Therefore, generality is not really restricted by considering only the case where P and Q are matrix polynomials in W .

3 The ODE Method

On-line network learning algorithms are of the general form

$$\theta_n = \Pi(\theta_{n-1} + \gamma_n h(z_n, \theta_{n-1})), \quad (10)$$

where θ is the vector of network parameters to be learned and θ_n is its estimate after n updating steps, z_n is the training pattern and γ_n the learning rate used at the n -th learning step, $h(\cdot, \cdot)$ is a function characteristic of the algorithm, and Π a “projection” mapping which may be necessary to keep the parameter updates constrained suitably.

The key tool in the analysis of the sequence $\{\theta_n\}$ is the so-called *interpolated process* $\theta^0(\cdot) = (\theta^0(t), t \geq 0)$, defined by

$$\theta^0(t) = \frac{t_n - t}{\gamma_n} \theta_{n-1} + \frac{t - t_{n-1}}{\gamma_n} \theta_n, \quad t_{n-1} \leq t < t_n, \quad (11)$$

where

$$t_0 = 0, \quad t_n = \gamma_1 + \dots + \gamma_n;$$

i.e., $\theta^0(\cdot)$ is obtained by piecewise constant interpolation of $\{\theta_n\}$ with interpolation intervals $\{\gamma_n\}$. Observe in particular that $\theta^0(t_n) = \theta_n$.

Kuan & Hornik (1990) investigate the properties of the interpolated process for the case of small constant learning rates and give applications to Error Back-Propagation in supervised learning and PCA feature extraction algorithms

based on feedforward architectures. In the present paper, we shall always assume that $\gamma_n \rightarrow 0$ at a suitable rate as $n \rightarrow \infty$. In this case, it can be shown that $\theta^0(\cdot)$ eventually follows the solution paths of an ODE (Ljung, 1977; Kushner & Clark, 1978; Ljung & Söderström, 1983). More precisely, Kushner’s method, which we find most convenient to use for our purpose, proceeds as follows.

Let us write m for the number of components of θ , and introduce left shifts $\theta^n(\cdot) = (\theta^n(t), t \geq 0)$ of the interpolated process (11) by means of $\theta^n(t) = \theta^0(t_n + t)$; observe that $\theta^n(0) = \theta^0(t_n) = \theta_n$. Clearly, all processes $\theta^n(\cdot)$ are elements of $C([0, \infty), \mathbb{R}^m)$, the space of all \mathbb{R}^m -valued continuous functions on $[0, \infty)$. Under suitable conditions, it can be shown that the set of processes $\{\theta^n(\cdot)\}$ is bounded and equicontinuous on $[0, T]$ for all $T < \infty$, such that by the famous Arzela-Ascoli Theorem (see e.g. Dunford & Schwartz, 1966), it is a *relatively compact* subset of $C([0, \infty), \mathbb{R}^m)$ if this space is given the topology of uniform convergence on bounded intervals. (I.e., for each infinite subsequence $\{n_l\}$ we can find a subsequence $\{n_l'\} \subset \{n_l\}$ such that $\theta^{n_l'}(\cdot)$ converges uniformly on bounded intervals.) If for the moment we assume that Π is just the identity mapping, then the limits of convergent subsequences satisfy the ODE $\dot{\theta} = \bar{h}(\theta)$, where the dot denotes the derivative with respect to t and \bar{h} is obtained from h by a suitable averaging procedure. More precisely, suppose that, as is the case for the feature extraction algorithms we are interested in, the learning patterns z_n can be decomposed as $z_n = (x_n, y_n)$, where the $\{x_n\}$ are, say, a sequence of independent random variables, and the y_n are generated throughout the algorithm according to

$$y_n = g(x_n, y_{n-1}, \theta_{n-1}) \quad (12)$$

with some initial y_0 , e.g. $y_0 = 0$. Hence, in general, y_n (implicitly) depends on all previous parameter updates $\theta_0, \dots, \theta_{n-1}$ and exogeneous network inputs x_1, \dots, x_n . For fixed θ , define a sequence $y_n(\theta)$ by means of the recursion

$$y_n(\theta) = g(x_n, y_{n-1}(\theta), \theta) \quad (13)$$

with initial condition $y_0(\theta) = y_0$. Then, provided that the limit exists and suitable additional assumptions are satisfied, we may take

$$\bar{h}(\theta) = \lim_{n \rightarrow \infty} \mathbf{E} h(x_n, y_n(\theta), \theta). \quad (14)$$

Now let Θ^* be the set of all asymptotically stable equilibria of the ODE $\dot{\theta} = \bar{h}(\theta)$ and $\mathcal{D}(\Theta^*)$ its domain of attraction. If one can show that $\{\theta_n\}$ enters some compact subset of $\mathcal{D}(\Theta^*)$ infinitely often, then the above approach allows

to conclude that θ_n converges to some $\theta^* \in \Theta^*$ as $n \rightarrow \infty$. However, the verification of this condition, leading to a global asymptotic analysis of the solutions of a usually complicated, nonlinear system of differential equations, unfortunately appears to be virtually impossible for many applications and in particular, for most feature extraction algorithms we are concerned with, the only exception that we are currently aware of being Oja’s one-unit algorithm which was fully analyzed in Oja & Karhunen (1985). In any case, the above characterization of the asymptotic paths of the interpolated process implies that for a “good” algorithm, the asymptotically stable equilibria, being at least locally attractive limit points of the associated ODE, should be desired limit points of the algorithm (i.e. points one actually wants the algorithm to converge to). On the other hand, if none of the desired limit points is an asymptotically stable equilibrium of the ODE, then we expect the performance of the algorithm to be rather poor. Therefore, an explicit characterization of Θ^* is of fundamental importance in understanding the asymptotic properties of the algorithm, even if one does not succeed in identifying $\mathcal{D}(\Theta^*)$.

For applicability of Kushner’s ODE method as outlined above, as well as for reasonable behavior of the algorithm, it is necessary that the sequence $\{\theta_n\}$ of parameter updates remains bounded or constrained to a suitable compact subset of \mathbb{R}^m . For example, we already know that if we use Földiák’s architecture with symmetric decorrelation, then the eigenvalues of W_n should eventually be less than one in absolute value. This goal can be accomplished by implementing some projection device Π as in equation (10). Such a device could e.g. truncate the entries of θ if they become too large; for feature extraction networks, this would enforce biologically very plausible limits to maximal interconnection strengths. Or, Π could project into a lower dimensional compact subset of \mathbb{R}^m , as is the case if the A part is trained using hebbian learning with rowwise euclidean normalization. As a rule of thumb, the ODE then becomes $\dot{\theta} \in D\Pi(\bar{h}(\theta))$, where $D\Pi$ is the set of all directional derivatives of Π ; for more details, see e.g. chapter 5.3 in Kushner & Clark (1978). In particular, if truncation is employed to constrain the updates to some hyperrectangle, then we still have $\dot{\theta} = \bar{h}(\theta)$ in the interior of the hyperrectangle.

4 Results

Now let us apply Kushner’s ODE method to the general class of PCA feature extraction algorithms introduced in section 2. In this case, $\theta = [\text{vec}(A)', \text{vec}(W)']'$, where the “vec” operator stacks one column above the other. Observe that ba-

sically all algorithms of interest keep W constrained to a lower dimensional subspace \mathcal{W} of the space of all $p \times p$ matrices. Hence, in a nonredundant parametrization $\theta \mapsto (A, W)$, θ contains the entries of A and the coordinates of W with respect to a suitable basis of \mathcal{W} . In the sequel, we shall find it notationally convenient to continue the analysis in terms of A and W , keeping in mind that $W \in \mathcal{W}$.

For sake of simplicity, let the learning rates be the same for both the A and the W part of the algorithm. The updating equations are

$$\Delta A_n = \gamma_n (y_n x_n' - \Phi(y_n y_n') A_{n-1}), \quad (15)$$

$$\Delta W_n = \gamma_n \Omega(y_n y_n'), \quad (16)$$

where the sequence $\{y_n\}$ is generated by $y_0 = 0$ and

$$y_n = P(W_{n-1})y_{n-1} + Q(W_{n-1})A_{n-1}x_n, \quad (17)$$

such that

$$\begin{aligned} y_n(\theta) &= P(W)y_{n-1}(\theta) + Q(W)Ax_n \\ &\vdots \\ &= \sum_{i=0}^{n-1} P(W)^i Q(W)Ax_{n-i}. \end{aligned}$$

For what follows, we assume that

[A 1] $\{x_n\}$ is a sequence of independent, bounded random vectors with mean zero and the same covariance matrix Σ .

[A 2] $\{\gamma_n\}$ is a sequence of positive numbers satisfying

$$\sum_n \gamma_n = \infty, \quad \sum_n \gamma_n^2 < \infty.$$

Using [A 1], we have

$$\mathbf{E}y_n(\theta)x_n' = Q(W)A\Sigma$$

independently of n , and the covariance matrix of $y_n(\theta)$ is given by

$$\mathbf{E}y_n(\theta)y_n(\theta)' = \sum_{i=0}^{n-1} P(W)^i Q(W)A\Sigma A' Q(W)' P(W)^i.$$

Hence, in order to guarantee that the limiting covariance matrix of $y_n(\theta)$ exists, we need all eigenvalues of $P(W)$ to be less than one in absolute value, in which case

$$\lim_{n \rightarrow \infty} \mathbf{E} y_n(\theta) y_n(\theta)' = \sum_{i=0}^{\infty} P(W)^i Q(W) A \Sigma A' Q(W)' P(W)^i =: R(A, W)$$

and, by combining (14) with (15) and (16) and the above computations, we obtain the asymptotic ODE

$$\dot{A} = Q(W) A \Sigma - \Phi(R) A \quad (18)$$

$$\dot{W} = \Omega(R) \quad (19)$$

with $R = R(A, W)$.

As $P(O) = O$ by assumption, the above eigenvalue condition is automatically satisfied if Ω sets at least all on- and superdiagonal entries to zero such that W remains “at most” subdiagonal throughout the algorithm. In this case, we necessarily have $P(W)^i = O$ for $i \geq p$ such that $R(A, W)$ is the sum of only finitely many terms. The following key result will be proved in the appendix.

Theorem 1. *Let $\{A_n\}$ and $\{W_n\}$ be generated by (15) and (16), respectively, and let $A^n(\cdot)$ and $W^n(\cdot)$ be the left shifts of the corresponding interpolated processes.*

Assume that \mathcal{W} contains only subdiagonal matrices and that with probability one, the sequence $\{A_n, W_n\}$ is bounded. Then, with probability one, $(A^n(\cdot), W^n(\cdot))$ is bounded and equicontinuous on bounded intervals. If $(A(\cdot), W(\cdot))$ is the limit of a convergent subsequence, it satisfies the ODE

$$\dot{A} = Q(W) A \Sigma - \Phi(R) A$$

$$\dot{W} = \Omega(R).$$

with $R = R(A, W)$ and $W \in \mathcal{W}$.

Let Θ^ be the set of all locally asymptotically stable equilibria (A, W) of the above ODE and $\mathcal{D}(\Theta^*)$ its domain of attraction. If (A_n, W_n) enters a compact subset S of $\mathcal{D}(\Theta^*)$ infinitely often with probability one, then $(A_n, W_n) \rightarrow \Theta^*$ with probability one.*

As already explained in section 3, boundedness of $\{A_n, W_n\}$ can always be ensured upon combination of the algorithm with a suitable projection mechanism which confines the updates to a bounded subset of the network weight

space. If the updates are constrained to some hypercube by a simple truncation device, the asymptotic paths satisfy the above ODE in the interior of that hypercube.

In the case where W is not constrained to subdiagonality, it has to be guaranteed that W_n eventually remains in the *stability region* which consists of all $W \in \mathcal{W}$ for which all eigenvalues of $P(W)$ are less than one in absolute value (i.e., the region where $R(A, W)$ is actually defined). Again, this can be ensured using simple truncation mechanisms. As an example, if $P(W) = W^k$ for some $k \geq 1$ as is the case in real-time applications based on Földiák's architecture, then it is easily seen that it suffices to keep $|\omega_{ij}| \leq p^{-1}$ for all off-diagonal entries ω_{ij} of W . To avoid unnecessary technicalities, we shall not formulate an explicit theorem and continue to refer to (18) and (19) as "the asymptotic ODE", although in general it will only be defined for W in a suitable neighborhood of O ; for more details, see Kushner & Clark (1978, p. 40 and section 5.3).

We now turn over to the investigation of the equilibria of the asymptotic ODE and their local stability properties. For this purpose, we shall for sake of simplicity make the additional assumption that

[A3] *All eigenvalues of Σ are distinct and positive.*

Let us write λ_i for the i -th largest eigenvalue of Σ and u_i for an associated unit length eigenvector which is then unique up to a change of sign. In what follows, it will also be convenient to let σ resp. τ denote the coefficient of W in $P(W)$ resp. $Q(W)$ such that

$$P(W) = \sigma W + \dots, \quad Q(W) = I + \tau W + \dots$$

where the dots indicate terms containing higher powers of W .

The equations for an equilibrium point of the ODE (18) and (19) are

$$Q(W)A\Sigma = \Phi(R)A, \quad \Omega(R) = O, \quad R = R(A, W). \quad (20)$$

Of course, an explicit solution of these equations is impossible without being more specific about the particular choice of Φ and Ω . Assume in addition that $Q(W)$ has full rank p ; in fact, this is the case for all relevant W in all applications of our interest. We then have the following general result.

Theorem 2. *Let A and W solve (20) such that $Q(W)$ has full rank p . Let $r := \text{rank}(A)$. Then there exist a $p \times r$ matrix C of full rank r and a sequence of mutually distinct indices $1 \leq i_1, \dots, i_r \leq d$ such that*

$$A = C[u_{i_1}, \dots, u_{i_r}]'. \quad (21)$$

For the proof of theorem 2, see the appendix. Geometrically speaking, equation (21) says that $\text{span}(A')$, the subspace of \mathbb{R}^d spanned by the rows of A , equals $\text{span}\{u_{i_1}, \dots, u_{i_r}\}$.

Of course, as we are interested in algorithms which extract the first p principal components, we expect any “reasonably good” algorithm to exhibit the following behavior. The set of asymptotically stable equilibria should not be empty, and all asymptotically stable equilibria should have $\text{rank}(A) = p$ and $\{i_1, \dots, i_p\} = \{1, \dots, p\}$; all other equilibria, corresponding to A 's which extract some “wrong” (or not enough) principal components, should be unstable. In addition to that, W should at least be subdiagonal to allow the mature network for finite-time exact computation of its output to a previously unseen input pattern. Hence in particular, if W is kept symmetric throughout the algorithm, all asymptotically stable equilibria should have $W = O$.

As will be shown in the appendix, equilibria with $\text{rank}(R(A, W)) < p$ are always unstable; hence, in the following discussion, we may restrict our attention to equilibria with full rank R .

For the local PCA algorithms which are the main concern of this paper, we have $\Phi = \text{diag}$ and $\Omega = -\text{offdiag}$ or $\Omega = -\text{subdiag}$; in the sequel, we shall refer to these two variants as a local PCA algorithm “in symmetric mode” respectively “in asymmetric mode”. In either case, due to the fact that $R(A, W)$ is symmetric, the equations for an equilibrium are

$$Q(W)A\Sigma = \text{diag}(R)A, \quad \text{offdiag}(R) = O, \quad R = R(A, W). \quad (22)$$

We have the following result.

Theorem 3. *All solutions of (22) with $W = O$ and full rank R are such that the rows of A are mutually perpendicular unit length eigenvectors of Σ , with associated (diagonal) eigenvalue matrix $R = A\Sigma A'$. If $\tau \neq 0$, these are the only equilibrium points with subdiagonal W and full rank R .*

Remark. If $\tau = 0$, there may be critical points with subdiagonal, but nonzero W and full rank R , see example 1 of the appendix. In the symmetric cases, a complete description of the set of equilibrium points (even with full rank R) is very hard. Example 2 of the appendix shows that each of the “desired” equilibria may actually lie on a curve of equilibria.

The local stability properties of the local PCA algorithms are given in the following two theorems. To simplify matters, let $\mathcal{J} = \{(i, j) : 1 \leq j < i \leq p\}$,

$$\mathcal{S} = \{(\sigma, \tau) : \sigma + \tau \geq 1, \sigma \leq 0\},$$

and let

$$\alpha_{ij}(\sigma, \tau) = \lambda_i + (\sigma + \tau - 1)\lambda_j, \quad \beta_{ij}(\sigma, \tau) = (\sigma + \tau)\lambda_i - \sigma\lambda_j.$$

Theorem 4. *For the local PCA algorithms in asymmetric mode, equilibria with $W = O$ are asymptotically stable if and only if*

$$A = [\pm u_1, \dots, \pm u_p]'$$

and

$$\alpha_{ij}(\sigma, \tau) > 0, \quad \beta_{ij}(\sigma, \tau) > 0 \quad \text{for all } (i, j) \in \mathcal{J}.$$

This condition is satisfied for arbitrary $\lambda_1 > \dots > \lambda_p > 0$ if and only if $(\sigma, \tau) \in \mathcal{S}$. If for some i , the i -th row of A does not equal $\pm u_i$, or if $\alpha_{ij}(\sigma, \tau) < 0$ or $\beta_{ij}(\sigma, \tau) < 0$ for some $(i, j) \in \mathcal{J}$, the equilibrium is unstable. This is always the case if $\sigma + \tau \leq 0$.

Corollary. *If $(\sigma, \tau) \in \mathcal{S}$, the asymptotically stable equilibria of the local PCA algorithms in asymmetric mode are exactly*

$$A = [\pm u_1, \dots, \pm u_p]', \quad W = O,$$

and all other equilibria are unstable.

For the algorithm of Rubner & Tavian (1990), $P(W) \equiv O$ and $Q(W) = I + W$, hence $(\sigma, \tau) = (0, 1) \in \mathcal{S}$ and the conclusions of the corollary apply. Algorithms based on the hierarchical modification of Földiák's architecture which perform k y -updating cycles use $P(W) = W^k$ and $Q(W) = I + W + \dots + W^{k-1}$. Thus, if $k = 1$, we have $\sigma = 1$ and $\tau = 0$ and all equilibria are unstable because for $i > j$,

$$\beta_{ij}(1, 0) = \lambda_i - \lambda_j < 0.$$

If $k > 1$, $(\sigma, \tau) = (0, 1) \in \mathcal{S}$, and again, the conclusions of the corollary apply. We infer that during the learning period, it is not necessary to perform exact decorrelation before updating the network weights, thereby motivating the use of "quicker" feedback algorithms with $1 < k < p$.

Theorem 5. *For a local PCA algorithm in symmetric mode, it is necessary for an equilibrium with $W = O$ to be asymptotically stable that $A = [\pm u_{i_1}, \dots, \pm u_{i_p}]'$ with $\{i_1, \dots, i_p\} = \{1, \dots, p\}$, and $\sigma < 0$. If $\sigma > 0$, all equilibria with $W = O$ are unstable.*

As real-time implementations of Földiák’s original algorithm use $P(W) = W^k$ for some $k \geq 1$ and therefore have $\sigma \geq 0$, we conclude that for these algorithms, none of the desired limit points is asymptotically stable. Similarly, as $\sigma = 0$ if $P(W) \equiv O$, we infer that no feedback-free local PCA algorithm gives rise to asymptotically stable desired equilibria if run in symmetric mode.

Therefore, hierarchical decorrelation should always be preferred over more competitive, symmetric decorrelation mechanisms, for reasons of superior performance of the algorithms — in symmetric mode, the desired limit points are not attractive enough. In fact, it is not too hard to give a biologically plausible interpretation of these findings. In our cases, the local updating of the A part of the network weights forces the rows of A to be eigenvectors of the input covariance matrix Σ ; in addition, if the network wants to extract as much input information as possible, the rows must not be collinear and hence they have to be mutually perpendicular. Thus, the local structure forces the units to follow a very strict hierarchy (of output variance). On the other hand, if each unit tries to maximize its output variance (subject to identical constraints) and all units are allowed to compete equally, then it will take much longer for a hierarchical structure to evolve than if this hierarchy is explicitly forced (or at least strongly supported) by the network interconnection topology. Of course, allowing for competition in order to obtain more “balanced” representations is quite attractive; however, this balance should be structurally stable.

For the subspace algorithm (1), Φ is the identity mapping and $\Omega \equiv 0$ such that $W \equiv O$. It is shown in Baldi & Hornik (1991) that all equilibria with full rank R are of the form $A = C[u_{i_1}, \dots, u_{i_p}]'$ where the i_j are mutually distinct and C is an arbitrary, orthogonal $p \times p$ matrix. Therefore, these equilibria are not isolated and thus cannot be asymptotically stable. This deficiency of the subspace algorithm was already noticed implicitly in Williams (1985). It can be shown that the component of small perturbations about an equilibrium which are perpendicular to $\text{span}(A')$ die out asymptotically if and only if $\{i_1, \dots, i_p\} = \{1, \dots, p\}$; a proof of this fact, together with a discussion of the evolution of the components along $\text{span}(A')$, see Krogh & Hertz (1990).

For Sanger’s GHA (2), $\Phi = \text{lower}$ and $\Omega \equiv 0$ such that $W \equiv O$. It is easily seen by induction (cf. e.g. the proof of theorem 3) that all equilibria with full rank R are of the form $A = [\pm u_{i_1}, \dots, \pm u_{i_p}]'$ with mutually distinct $1 \leq i_1, \dots, i_p \leq d$. The following result is implicitly contained in both Oja & Karhunen (1985) and Sanger (1989).

Theorem 6. *The asymptotically stable equilibria of Sanger’s GHA (2) are*

$$A = [\pm u_1, \dots, \pm u_p]';$$

all other equilibria are unstable.

5 Remarks and Problems

In Oja & Karhunen (1985), it is shown (lemma 5, p. 80) that if Oja’s one-unit algorithm is used with uniformly bounded inputs and the learning rates are sufficiently small (but do not necessarily tend to 0), then the weight updates automatically remain inside some bounded subset of the weight space. It is definitely worthwhile investigating whether or not multioutput generalizations also possess this very strong stability property, in particular if local algorithms with asymmetric decorrelation do so. If this were the case, explicit truncation would become obsolete. However, this question appears to be extremely challenging.

If the ODE method is to be used for producing explicit convergence results for on-line PCA learning algorithms, $\mathcal{D}(\Theta^*)$, the domain of attraction of the set of asymptotically stable equilibria, has to be identified (or at least, one should succeed in exhibiting a suitably “rich” subset of $\mathcal{D}(\Theta^*)$ which the updates could be confined to). However, as already indicated in section 3, the asymptotic ODEs of multioutput algorithms are quite complicated, thereby making such an identification very hard.

In fact, the only case where a complete global analysis of the ODE is available appears to be Oja’s one-unit algorithm (Oja & Karhunen, 1985). For the subspace algorithm, the analysis in both Williams (1985) and Krogh & Hertz (1990) is strictly local, and it actually seems to still be unknown whether an equilibrium of the form $A = [\pm u_1, \dots, \pm u_p]'$ is stable or unstable (remember that it cannot be asymptotically stable). Sanger (1989, p. 463) claims that for the GHA, “the domain of attraction (of $\Theta^* = \{A : A = [\pm u_1, \dots, \pm u_p]'\}$) includes all matrices with bounded weights”, which is obviously wrong due to the existence of unstable equilibria of the form $A = [\pm u_{i_1}, \dots, \pm u_{i_p}]'$ with $(i_1, \dots, i_p) \neq (1, \dots, p)$. Clearly, for local PCA algorithms the global asymptotic analysis is even harder. However, exhaustive computer simulations confirm rapid convergence to Θ^* in asymmetric mode even when starting with very large weights or initial configurations very close to unstable equilibria.

Finally, let us mention that the asymptotic analysis for local PCA algorithms which update A using hebbian learning with explicit rowwise euclidean

normalization rather than apply Oja's one-unit algorithm to each of the rows of A is more or less identical to the analysis presented here, cf. e.g. the results in Oja & Karhunen (1985).

Appendix: Mathematical Proofs

Proof of theorem 1. Let Θ_s be the set of all $\theta = [\text{vec}(A)', \text{vec}(W)']'$, where A is an arbitrary $p \times d$ matrix and W is a subdiagonal $p \times p$ matrix. We proceed by applying theorem 2.5.2 of Kushner & Clark (1978) with the obvious modification that as by assumption the algorithm keeps θ_n in Θ_s , all conditions have to be verified only for θ in Θ_s . (Alternatively, one could, at the expense of more complicated notations, work with a nonredundant parametrization $\theta \mapsto (A, W)$.)

In vectorized form, the algorithm is

$$\theta_n = \theta_{n-1} + \gamma_n h(z_n, \theta_{n-1}),$$

where $z_n = (x_n, y_n)$,

$$h(z, \theta) = \begin{bmatrix} \text{vec}(yx' - \Phi(yy')A) \\ \text{vec}(\Omega(yy')) \end{bmatrix}$$

and the y_n are generated according to (17). We already know that for all $\theta \in \Theta_s$,

$$\lim_{n \rightarrow \infty} \mathbf{E} h(z_n(\theta), \theta) = \begin{bmatrix} \text{vec}(Q(W)A\Sigma - \Phi(R)A) \\ \text{vec}(\Omega(R)) \end{bmatrix} := \bar{h}(\theta), \quad (23)$$

where of course $R = R(A, W)$.

Applied to our case, theorem 2.5.2 of Kushner & Clark (1978) states that if conditions [A 2.2.3], [A 2.4.5], [A 2.5.2] and [A 2.5.3] of Kushner & Clark (1978) are satisfied and $\{\theta_n\}$ remains bounded with probability one, then $\{\theta^n(\cdot)\}$ is bounded and equicontinuous on bounded intervals with probability one, and the limits of convergent subsequences satisfy the ODE $\dot{\theta} = \bar{h}(\theta)$. As we assumed the boundedness condition, the proof of theorem 1 is completed upon verification of the abovementioned conditions.

[A 2.2.3] is trivially satisfied. We start with [A 2.5.2].

A 2.5.2. *There is a continuous function $\bar{h}(\cdot)$ such that for some $T > 0$, for each $\epsilon > 0$ and each $\theta \in \Theta_s$,*

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i (h(z_i(\theta), \theta)) - \bar{h}(\theta)) \right| \geq \epsilon \right\} = 0,$$

where

$$m(t) = \max\{n : t_n \leq t\}.$$

If $\bar{h}(\cdot)$ is defined by (23), it is clearly continuous for $\theta \in \Theta$. Fix $\theta \in \Theta$, and $T > 0$, and let $h_n(\theta) = h(z_n(\theta), \theta)$. Observe that for $t \leq T$,

$$\sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i \leq T$$

for all j . For each $\epsilon > 0$, we can choose n sufficiently large such that for all $i \geq m(jT)$ and $j \geq n$, $|\mathbf{E} h_i(\theta) - \bar{h}(\theta)| \leq \epsilon/(2T)$. Hence for n sufficiently large,

$$\begin{aligned} & \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i (h_i(\theta) - \bar{h}(\theta)) \right| \\ & \leq \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i (h_i(\theta) - \mathbf{E} h_i(\theta)) \right| \\ & \quad + \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i (\mathbf{E} h_i(\theta) - \bar{h}(\theta)) \right| \\ & \leq \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i (h_i(\theta) - \mathbf{E} h_i(\theta)) \right| + \epsilon/2. \end{aligned}$$

Note that almost sure convergence of $\sum_{i=1}^n \gamma_i (h_i(\theta) - \mathbf{E} h_i(\theta))$ is equivalent to the condition that for each $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \sup_{m \geq n} \left| \sum_{i=n}^m \gamma_i (h_i(\theta) - \mathbf{E} h_i(\theta)) \right| < \epsilon/2 \right\} = 1,$$

which in turn implies that

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i (h_i(\theta) - \mathbf{E} h_i(\theta)) \right| < \epsilon/2 \right\} = 1.$$

Hence [A 2.5.2] is established if we show that $\sum_{i=1}^n \gamma_i (h_i(\theta) - \mathbf{E} h_i(\theta))$ converges with probability one, which can be done by showing that $\{\gamma_n(h_n(\theta) - \mathbf{E} h_n(\theta))\}$ is a mixingale of size $-1/2$ (or larger) with square summable magnitude indices and using the mixingale convergence theorem (McLeish, 1975, corollary 1.8).

As $\{x_n\}$ is bounded by assumption, $\{y_n(\theta)\}$ is clearly bounded as well. By lemma 1 in Andrews (1989), $\text{vec}(y_n(\theta)x_n')$ and $\text{vec}(y_n(\theta)y_n(\theta)')$ are near epoch dependent (NED, cf. Gallant & White, 1988) on $\{x_n\}$ of arbitrarily large size.

It follows by lemma 3.14 of Gallant & White (1988) that $\{h_n(\theta) - \mathbf{E} h_n(\theta)\}$ is a bounded mixingale of arbitrarily large size and with bounded magnitude indices $\{c_n\}$. Hence, $\{\gamma_n(h_n(\theta) - \mathbf{E} h_n(\theta))\}$ is also a bounded mixingale of arbitrarily large size and magnitude indices $\{\gamma_n c_n\}$ which is square summable by assumption [A 2] and boundedness of $\{c_n\}$. Therefore, the mixingale convergence theorem applies, and [A 2.5.2] is established.

If a sequence $\{y_n\}$ is generated by $y_n = P_{n-1}y_{n-1} + c_n$, we find upon resubstitution that for all k ,

$$y_n = L_{n-1,k}y_{n-k} + \sum_{i=0}^{k-1} L_{n-1,i}c_{n-i},$$

where $L_{n-1,0} = I$ and for $i \geq 1$, $L_{n-1,k} = P_{n-1} \cdots P_{n-k}$. In particular, if all P_n are subdiagonal $p \times p$ matrices, $L_{n,k} = O$ for all $k \geq p$, and

$$y_n = \sum_{i=0}^{p-1} L_{n-1,i}c_{n-i}. \quad (24)$$

Now let $P_n = P(W_n)$, $c_n = Q(W_{n-1})A_{n-1}x_n$ with $\theta_n = [\text{vec}(A_n)', \text{vec}(W_n)']'$ a bounded sequence in Θ_s . Using the above formula, it is readily seen that for all $\epsilon > 0$ we can find $\delta > 0$ such that if $\{\tilde{\theta}_n\}$ is another sequence in Θ_s satisfying $\max_{m-k \leq n \leq m+l} |\theta_n - \tilde{\theta}_n| < \delta$ and \tilde{y}_n is generated by $\tilde{y}_n = P(\tilde{W}_{n-1})\tilde{y}_{n-1} + Q(\tilde{W}_{n-1})\tilde{A}_{n-1}x_n$, then $\max_{m \leq n \leq m+l} |y_n - \tilde{y}_n| < \epsilon$, which in turn implies [A 2.5.3].

Let \otimes denote the Kronecker product of two matrices (see e.g. Magnus & Neudecker, 1988) such that $\text{vec}(LMN) = (N' \otimes L)\text{vec}(M)$ for all matrices L , M , N of compatible dimensions. Then $\text{vec}(\Phi(yy')A) = (A' \otimes I)\text{vec}(\Phi(yy'))$, and thus

$$|h(z, \theta)| \leq \underbrace{|A' \otimes I|}_{\bar{g}_1(\theta)} \underbrace{|\text{vec}(\Phi(yy'))|}_{g_3(z)} + \underbrace{|\text{vec}(yx')| + |\text{vec}(\Omega(yy'))|}_{g_4(z)}.$$

Clearly, $\bar{g}_1(\cdot)$ is bounded on bounded subsets of Θ_s . For $M > 0$, let \mathcal{E}_M be the event that for some $l \in \{3, 4\}$ and some n , $g_l(z_n) > M$. As by assumption $\{\theta_n\}$ and $\{x_n\}$ are bounded with probability one, (24) shows that $\{y_n\}$ and hence also $\{g_l(z_n)\}$ remain bounded with probability one; hence $\lim_{M \rightarrow \infty} \mathbf{P}(\mathcal{E}_M) = 0$. On the complement of \mathcal{E}_M ,

$$\sup_{j \geq n} \max_{t \leq \Delta} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i g_l(z_i) \right| \leq M \sup_{j \geq n} \max_{t \leq \Delta} \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i \leq M\Delta.$$

Let $\epsilon > 0$ be arbitrary. As soon as $\Delta < \epsilon/M$,

$$\mathbf{P} \left\{ \sup_{j \geq n} \max_{t \leq \Delta} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i g_t(z_i) \right| \geq \epsilon \right\} \leq \mathbf{P}(\mathcal{E}_M).$$

Now let $M \rightarrow \infty$ to conclude that for all $\epsilon > 0$ and $l \in \{3, 4\}$,

$$\lim_{\Delta \rightarrow 0} \mathbf{P} \left\{ \sup_{j \geq n} \max_{t \leq \Delta} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i g_t(z_i) \right| \geq \epsilon \right\} = 0,$$

establishing [A 2.4.5] and thereby completing the proof of theorem 1.

Proof of theorem 2. Letting $M = Q(W)^{-1}\Phi(R)$, the first equation of (20) becomes $A\Sigma = MA$. Hence,

$$M Au_i = A\Sigma u_i = \lambda_i Au_i, \quad i = 1, \dots, d,$$

from which we conclude that Au_i is either zero or an eigenvector of M with eigenvalue λ_i . As by [A 3] all λ_i are distinct, the nonzero Au_i are linearly independent by a well-known result from linear algebra; on the other hand, the number of linearly independent (and thus nonzero) Au_i equals $\text{rank}(A[u_1, \dots, u_d]) = \text{rank}(A) = r$. Now let $U_0 = [u_{i_1}, \dots, u_{i_r}]D$ where D is a suitable diagonal matrix with entries ± 1 and i_1, \dots, i_r is some permutation of the indices in $\mathcal{I}_0 = \{i : Au_i \neq 0\}$; similarly, let j_1, \dots, j_{d-r} be the indices in $\mathcal{I}_\perp = \{i : Au_i = 0\}$ arranged in ascending order and $U_\perp = [u_{j_1}, \dots, u_{j_{d-r}}]$. Then clearly $C := AU_0D$ has full rank r , $AU_\perp = O$, such that finally, as u_i and u_j are perpendicular for $i \neq j$ and thus $I = U_0U_0' + U_\perp U_\perp'$, $A = A(U_0U_0' + U_\perp U_\perp') = AU_0U_0' = C[u_{i_1}, \dots, u_{i_r}]'$.

Proof of theorem 3. If $W = O$, the equations for a critical point give

$$A\Sigma = \text{diag}(R)A, \quad \text{offdiag}(R) = O, \quad R = A\Sigma A'.$$

If R is full rank, so is A . It follows that the rows of A are eigenvectors of Σ with corresponding eigenvalue matrix R , and that

$$A\Sigma = \text{diag}(A\Sigma A')A = A\Sigma A'.$$

Multiplying by A' from the right we finally conclude that $AA' = I$, whence the first assertion of the theorem.

The proof of the second assertion is more involved. Let e_i denote the i -th Cartesian unit vector, and let $a'_i = e'_i A$ be the i -th row of A so that $A' = [a_1, \dots, a_p]$. The equations for a critical point are $\text{offdiag}(R) = O$ and, taking transposes for notational convenience,

$$\Sigma A' Q(W)' = A' \text{diag}(R). \quad (25)$$

Observe that R is a nonnegative definite diagonal matrix which is full rank by assumption. Hence, all its diagonal entries $\rho_i = e'_i R e_i$ are strictly positive. If W is subdiagonal, we can write W' as

$$W' = \begin{bmatrix} 0 & \omega_{2,1} & \omega_{3,1} & \cdots & \omega_{p,1} \\ & 0 & \omega_{3,2} & \cdots & \omega_{p,2} \\ & & \ddots & & \vdots \\ & & & 0 & \omega_{p,p-1} \\ & & & & 0 \end{bmatrix}.$$

Trivially $W'e_1 = 0$, and thus $Q(W)'P(W)^m e_1$ equals e_1 for $m = 0$ and zero for $m > 0$. Hence,

$$\rho_1 = e'_1 R e_1 = e'_1 A \Sigma A' e_1 = a'_1 \Sigma a_1$$

and

$$\rho_1 a_1 = A' \text{diag}(R) e_1 = \Sigma A' Q(W)' e_1 = \Sigma A' e_1 = \Sigma a_1.$$

As $\rho_1 > 0$, we conclude that a_1 is a unit length eigenvector of Σ with corresponding eigenvalue ρ_1 .

We now proceed by induction. Suppose we have already shown that for all $i < l$, $W'e_i = 0$ and that the a_i are mutually perpendicular unit length eigenvectors of Σ with corresponding eigenvalues ρ_i . Then, if $i < l$, we have $Q(W)'e_i = e_i$ and $P(W)^m e_i = 0$ for $m > 0$; if both $i, j < l$,

$$a'_i \Sigma a_j = \rho_i a'_i a_j = \begin{cases} \rho_i, & i = j \\ 0, & i \neq j. \end{cases}$$

Observing that

$$W'e_l = \omega_{l,1} e_1 + \cdots + \omega_{l,l-1} e_{l-1},$$

we obtain

$$W'^2 e_l = W'(\omega_{l,1} e_1 + \cdots + \omega_{l,l-1} e_{l-1}) = \omega_{l,1} W' e_1 + \cdots + \omega_{l,l-1} W' e_{l-1} = 0$$

and thus also $W'^m e_l = 0$, $m \geq 2$. In particular, $Q(W)'e_l = (I + \tau W')e_l$.

Now let $i < l$. As R is diagonal,

$$0 = e_i' R e_l = e_i' A \Sigma A' Q(W)' e_l = a_i' \Sigma A' (I + \tau W') e_l = a_i' \Sigma a_l + \tau \rho_i \omega_{l,i}.$$

On the other hand, multiplication of equation (25) by a_i' from the left and by e_l from the right gives

$$a_i' \Sigma a_l + \tau \rho_i \omega_{l,i} = \rho_l a_i' a_l.$$

Combining both equations and using that ρ_l is positive by assumption, we conclude that for all $i < l$, $a_i' a_l = 0$. Thus, $\tau \rho_i \omega_{l,i} = -a_i' \Sigma a_l = -\tau \rho_i a_i' a_l = 0$, whence $\omega_{l,i} = 0$ because both τ and ρ_i are nonzero. Summing up, $W' e_l = 0$ and a_l is perpendicular to all a_i with $i < l$.

This finally gives

$$\rho_l = e_l' R e_l = a_l' \Sigma a_l$$

and

$$\rho_l a_l = A' \text{diag}(R) e_l = \Sigma A' Q(W)' e_l = \Sigma A' e_l = \Sigma a_l,$$

hence a_l is a unit length eigenvector of Σ with corresponding eigenvalue ρ_l , completing the induction step.

Example 1. As an example for the possible existence of equilibria with full rank R and subdiagonal, but nonzero W in the case where $\tau = 0$, consider $P(W) = W$, $Q(W) \equiv I$, and $p = 2$. (The example may trivially be extended to larger values of p .) Let us write $\omega = \omega_{21}$ for the only nonzero entry of W , let u be a unit length eigenvector of Σ with associated eigenvalue λ , and let $A = [u, 0]'$. It is easily seen that for all ω , we have $A \Sigma = R A$, where $R = R(A, W)$ is diagonal with entries λ and $\omega^2 \lambda$. Hence, whenever $\omega \neq 0$, (A, W) is an equilibrium with subdiagonal, nonzero W and full rank R . This example also shows that if $\tau = 0$, there may be equilibria with subdiagonal, nonzero W , full rank R , and rank deficient A .

Example 2. We now show that equilibria with $W = O$ can actually lie on a whole curve resp. surface of equilibria with symmetric W . Let $P(W) = W$, $Q(W) \equiv I$, and again for sake of notational simplicity let $p = 2$ (the example also works for $p > 2$). Consider an equilibrium with $W = O$ and $A = [u, v]'$, where u and v are mutually perpendicular eigenvectors of Σ with corresponding eigenvalues λ and μ . Whenever $\omega^2 \leq \min(\lambda/\mu, \mu/\lambda)$, we can find $\alpha = \alpha(\omega)$, $\beta = \beta(\omega)$ satisfying

$$\alpha^2 = 1 - \omega^2 \frac{\mu}{\lambda}, \quad \beta^2 = 1 - \omega^2 \frac{\lambda}{\mu}$$

such that

$$\begin{aligned}\alpha^2\lambda + \beta^2\mu &= (1 - \omega^2)(\lambda + \mu), \\ \alpha^2\lambda - \beta^2\mu &= (1 + \omega^2)(\lambda - \mu).\end{aligned}$$

Now let

$$A(\omega) = [\alpha(\omega)u, \beta(\omega)v]', \quad W(\omega) = \begin{bmatrix} 0 & \omega \\ \omega & 0 \end{bmatrix}.$$

Then clearly $A(\omega)\Sigma = \text{diag}(\lambda, \mu)A(\omega)$, and, as $W(\omega) = VD(\omega)V'$, where

$$D(\omega) = \text{diag}(\omega, -\omega), \quad V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

we have

$$\begin{aligned}R(A(\omega), W(\omega)) &= \sum_{i=0}^{\infty} W(\omega)^i A(\omega)\Sigma A(\omega)' W(\omega)^{i'} \\ &= \sum_{i=0}^{\infty} VD(\omega)^i V' \begin{bmatrix} \alpha(\omega)^2\lambda & \\ & \beta(\omega)^2\mu \end{bmatrix} VD(\omega)^i V' \\ &= \frac{1}{2} \sum_{i=0}^{\infty} VD(\omega)^i \begin{bmatrix} \alpha(\omega)^2\lambda + \beta(\omega)^2\mu & \alpha(\omega)^2\lambda - \beta(\omega)^2\mu \\ \alpha(\omega)^2\lambda - \beta(\omega)^2\mu & \alpha(\omega)^2\lambda + \beta(\omega)^2\mu \end{bmatrix} D(\omega)^i V' \\ &= \frac{1}{2} V \begin{bmatrix} (\alpha(\omega)^2\lambda + \beta(\omega)^2\mu)/(1 - \omega^2) & (\alpha(\omega)^2\lambda - \beta(\omega)^2\mu)/(1 + \omega^2) \\ (\alpha(\omega)^2\lambda - \beta(\omega)^2\mu)/(1 + \omega^2) & (\alpha(\omega)^2\lambda + \beta(\omega)^2\mu)/(1 - \omega^2) \end{bmatrix} V' \\ &= \frac{1}{2} V \begin{bmatrix} \lambda + \mu & \lambda - \mu \\ \lambda - \mu & \lambda + \mu \end{bmatrix} V' \\ &= \begin{bmatrix} \lambda & \\ & \mu \end{bmatrix}.\end{aligned}$$

Hence, whenever $\omega^2 \leq \min(\lambda/\mu, \mu/\lambda)$, $A(\omega)$ and $W(\omega)$ satisfy (22). As ω varies, we obtain a curve of distinct equilibria, which for $\omega = 0$ contains the desired equilibrium A, W .

Stability analysis of the equilibria. For the remaining proofs of theorems 4 to 6, it will be convenient to start the analysis of the local stability properties of the equilibria at a general level which in fact pertains to all PCA algorithms considered in this paper. We shall continue to use the notations introduced in the proof of theorem 2; in addition to that, let $C_0 = AU_0$ and let Λ_0 and Λ_{\perp} denote the diagonal matrices with entries $\lambda_{i_1}, \dots, \lambda_{i_r}$ respectively $\lambda_{j_1}, \dots, \lambda_{j_{d-r}}$ such that $\Sigma U_0 = U_0 \Lambda_0$ and $\Sigma U_{\perp} = U_{\perp} \Lambda_{\perp}$.

To investigate the local stability properties of an equilibrium, we have to consider the evolution of small perturbations about the equilibrium. Let E and H be the perturbations of A and W , respectively. After linearization we obtain

$$\begin{aligned}\dot{E} &= Q(W)E\Sigma + dQ(H;W)A\Sigma - \Phi(R)E - \Phi(dR(E, H; A, W))A \quad (26) \\ \dot{H} &= \Omega(dR(E, H; A, W)), \quad (27)\end{aligned}$$

with the additional constraint $H \in \mathcal{W}$; here, $dQ(H;W)$ etc. denote Fréchet differentials, i.e.

$$Q(W + \epsilon H) = Q(W) + \epsilon dQ(H;W) + O(\epsilon^2) \quad \text{as } \epsilon \rightarrow 0.$$

Now notice that dR depends on E only via $E\Sigma A' = E\Sigma U_0 C_0' = EU_0 \Lambda_0 C_0'$ and its transpose; hence, $dR(E, H; A, W)$ is of the form $\Upsilon(EU_0, H; A, W)$. Let $E_0 = EU_0$ and $E_\perp = EU_\perp$. As $A\Sigma U_\perp = AU_\perp \Lambda_\perp = O$, (26) and (27) are equivalent to

$$\dot{E}_\perp = Q(W)E_\perp \Lambda_\perp - \Phi(R)E_\perp, \quad (28)$$

$$\begin{aligned}\dot{E}_0 &= Q(W)E_0 \Lambda_0 + dQ(H;W)C_0 \Lambda_0 \\ &\quad - \Phi(R)E_0 - \Phi(\Upsilon(E_0, H; A, W))C_0, \quad (29)\end{aligned}$$

$$\dot{H} = \Omega(\Upsilon(E_0, H; A, W)). \quad (30)$$

Equations (28) resp. (29) describe the evolution of perturbations of A perpendicular to resp. along $\text{span}(A')$, expressed in terms of the bases given by the columns of U_\perp resp. U_0 . Obviously, equation (28) does not depend upon E_0 and H , whereas (29) and (30) are independent from E_\perp .

Writing $v_\perp = \text{vec}(E_\perp)$, (28) becomes

$$\dot{v}_\perp = (\Lambda_\perp \otimes Q(W) - I \otimes \Phi(R))v_\perp$$

and we immediately deduce the following result.

Lemma. *Let A, W be an equilibrium of the asymptotic ODE. Then small perturbations of A perpendicular to $\text{span}(A')$ die out asymptotically if and only if all eigenvalues of the matrix*

$$K(A, W) = \Lambda_\perp \otimes Q(W) - I \otimes \text{diag}(R)$$

have negative real parts.

Hence, an equilibrium A, W can only be asymptotically stable if the above condition is satisfied; in particular, R has to be full rank. If one of the eigenvalues of $K(A, W)$ has positive real part, the equilibrium is unstable.

In all subsequent proofs, equilibria with $W = O$ are of interest. In these cases, as $R(C_0 U'_0, O) = C_0 \Lambda_0 C'_0$, $dQ(H, O) = \tau H$ and

$$dR(E, H; C U'_0, O) = (\sigma + \tau) H R + E_0 \Lambda_0 C'_0 + C_0 \Lambda_0 E'_0 + (\sigma + \tau) R H',$$

(29) and (30) become

$$\begin{aligned} \dot{E}_0 &= E_0 \Lambda_0 + \tau H C_0 \Lambda_0 - \Phi(R) E_0 \\ &\quad - \Phi((\sigma + \tau) H R + E_0 \Lambda_0 C'_0 + C_0 \Lambda_0 E'_0 + (\sigma + \tau) R H') C_0 \end{aligned} \quad (31)$$

$$\dot{H} = \Omega((\sigma + \tau) H R + E_0 \Lambda_0 C'_0 + C_0 \Lambda_0 E'_0 + (\sigma + \tau) R H') \quad (32)$$

with $H \in \mathcal{W}$.

Proof of theorems 4/5. If R is not full rank, the above lemma yields that the equilibrium is unstable. Hence, suppose that R is full rank. Using the first assertion of theorem 3, we infer that we may take $U_0 = A$ such that $\Lambda_0 = R$ and $C_0 = I$. As the eigenvalues of $K(U_0, O) = \Lambda_\perp \otimes I - I \otimes \Lambda_0$ are $\lambda_j - \lambda_i$ with $(i, j) \in \mathcal{I}_0 \times \mathcal{I}_\perp$, the lemma implies that the equilibrium can only be asymptotically stable if $\mathcal{I}_0 = \{1, \dots, p\}$, and is unstable otherwise.

As $R = \Lambda_0$ is diagonal and H has zero diagonal, we find that $\text{diag}(H R) = \text{diag}(R H') = O$ such that (31) and (32) simplify to

$$\dot{E}_0 = E_0 R + \tau H R - R E_0 - 2 \text{diag}(E_0) R. \quad (33)$$

$$\dot{H} = \Omega((\sigma + \tau) H R + E_0 R + R E'_0 + (\sigma + \tau) R H'), \quad (34)$$

where of course $\Omega = -\text{offdiag}$ in symmetric mode and $\Omega = -\text{subdiag}$ in asymmetric mode. The proof can now easily be completed by noticing that (33) and (34) can be decomposed into independent, one- resp. three-dimensional sub-problems. In fact, let ξ_{ij} and η_{ij} be the (i, j) -th component of E_0 resp. H , and denote the i -th diagonal entry of R by ρ_i . For $i = j$, we have $\eta_{ii} \equiv 0$ and

$$\dot{\xi}_{ii} = -2 \rho_i \xi_{ii},$$

which tends to zero exponentially fast.

Now let $i > j$. In symmetric mode, we have $\eta_{ij} \equiv \eta_{ji}$ and we obtain the systems

$$\dot{v}_{ij} = M_{ij}^s(\sigma, \tau) v_{ij},$$

where

$$v_{ij} = \begin{bmatrix} \xi_{ij} \\ \eta_{ij} \\ \xi_{ji} \end{bmatrix}, \quad M_{ij}^s(\sigma, \tau) = \begin{bmatrix} \rho_j - \rho_i & \tau \rho_j & \\ -\rho_j & -(\sigma + \tau)(\rho_i + \rho_j) & -\rho_i \\ & \tau \rho_i & \rho_i - \rho_j \end{bmatrix}.$$

It is straightforwardly computed that

$$\det(M_{ij}^s(\sigma, \tau)) = \sigma(\rho_i - \rho_j)^2(\rho_i + \rho_j).$$

As the degree of the characteristic polynomial $\phi_{ij}^s(\lambda) = \det(M_{ij}^s(\sigma, \tau) - \lambda I)$ is 3, its roots, being the eigenvalues of $M_{ij}^s(\sigma, \tau)$, are either all real, or one is real and two are complex conjugate. Hence, if, as has to be the case for asymptotic stability, all eigenvalues have negative real parts, the determinant has to be negative, which is only possible if $\sigma < 0$. On the other hand, if $\sigma > 0$, then the determinant is positive, hence at least one of the eigenvalues has positive real part and the equilibrium is unstable, thereby completing the proof of theorem 5.

In asymmetric mode, we have $\eta_{ji} \equiv 0$ (remember that $i > j$) and we obtain the systems

$$\dot{v}_{ij} = M_{ij}^a(\sigma, \tau) v_{ij},$$

where now

$$M_{ij}^a(\sigma, \tau) = \begin{bmatrix} \rho_j - \rho_i & \tau \rho_j & \\ -\rho_j & -(\sigma + \tau)(\rho_i + \rho_j) & -\rho_i \\ & & \rho_i - \rho_j \end{bmatrix}.$$

Hence, one eigenvalue of $M_{ij}^a(\sigma, \tau)$ is $\rho_i - \rho_j$, from which we conclude that the equilibrium can only be asymptotically stable if the ρ_i are arranged in descending magnitude, which together with the condition $\mathcal{I}_0 = \{1, \dots, p\}$ we already established implies that $\rho_1 = \lambda_1, \dots, \rho_p = \lambda_p$ and $A = [\pm u_1, \dots, \pm u_p]'$, and that otherwise it is unstable. In the case where $\rho_1 = \lambda_1, \dots, \rho_p = \lambda_p$, the remaining eigenvalues are the roots of the equations

$$\lambda^2 + \underbrace{(\lambda_i - \lambda_j + (\sigma + \tau)\lambda_j)}_{= \alpha_{ij}(\sigma, \tau)} \lambda + \underbrace{(\sigma + \tau)\lambda_j(\lambda_i - \lambda_j) + \tau\lambda_j^2}_{= \lambda_j \beta_{ij}(\sigma, \tau)} = 0.$$

It is easily seen that both roots of the quadratic polynomial $\lambda^2 + \alpha\lambda + \beta$ have negative real parts if and only if both $\alpha > 0$ and $\beta > 0$, and that at least one root has positive real part if $\alpha < 0$ or $\beta < 0$. Applied to our case, we thus have asymptotic stability if and only if for all $(i, j) \in \mathcal{J}$ we have $\alpha_{ij}(\sigma, \tau) > 0$ and $\beta_{ij}(\sigma, \tau) > 0$, and conversely, if for some $(i, j) \in \mathcal{J}$ we have $\alpha_{ij}(\sigma, \tau) < 0$ or $\beta_{ij}(\sigma, \tau) < 0$, the equilibrium is unstable. In particular, this is the case if $\sigma + \tau \leq 0$, because then $\beta_{ij}(\sigma, \tau) \leq \lambda_i - \lambda_j < 0$.

Finally, it is clear that the inequalities

$$\lambda_i + (\sigma + \tau - 1)\lambda_j > 0, \quad (\sigma + \tau)\lambda_i - \sigma\lambda_j > 0$$

can be valid for arbitrary $\lambda_1 > \dots > \lambda_p > 0$ and $(i, j) \in \mathcal{J}$ if and only if $\sigma + \tau \geq 1$ and $\sigma \leq 0$, i.e. if and only if $(\sigma, \tau) \in \mathcal{S}$, and the proof of theorem 4 is complete.

Proof of the corollary. If W is subdiagonal, $Q(W)$ is lower diagonal with $\text{diag}(Q(W)) = I$ and hence full rank. By our lemma, all equilibria with rank deficient R are unstable. Now observe that $(\sigma, \tau) \in \mathcal{S}$ implies in particular that $\tau \geq 1$, hence by theorem 3 all equilibria with full rank R are of the form $A = [\pm u_{i_1}, \dots, \pm u_{i_p}]'$ and $W = O$. By theorem 4, these equilibria are asymptotically stable if and only if $i_1 = 1, \dots, i_p = p$, and unstable otherwise, whence the corollary.

Proof of theorem 6. For Sanger's GHA, $W \equiv O$ and hence $Q(W) \equiv I$. Again using the lemma, all equilibria with $\text{rank}(R) < p$ are unstable. The equilibria with full rank R can be shown to be $A = [\pm u_{i_1}, \dots, \pm u_{i_p}]'$ such that again we may take $U_0 = A$, $\Lambda_0 = R$, and $C_0 = I$, and by applying the lemma once more it follows that the equilibrium is unstable if $\mathcal{I}_0 \neq \{1, \dots, p\}$. Trivially $H \equiv O$, and (31) becomes

$$\dot{E}_0 = E_0 R - R E_0 - \text{lower}(E_0 R + R E_0')$$

which now decomposes into one- and two-dimensional subsystems. For $i = j$, we have

$$\dot{\xi}_{ii} = -2 \rho_i \xi_{ii}$$

which tends to zero exponentially fast, and for $i > j$ we obtain

$$\begin{bmatrix} \dot{\xi}_{ij} \\ \dot{\xi}_{ji} \end{bmatrix} = \begin{bmatrix} -\rho_i & -\rho_i \\ & \rho_i - \rho_j \end{bmatrix} \begin{bmatrix} \xi_{ij} \\ \xi_{ji} \end{bmatrix}$$

and we immediately deduce that the equilibrium is asymptotically stable if and only if $i_1 = 1, \dots, i_p = p$, and otherwise unstable.

References

- Andrews, D. W. K. (1989). *An empirical process central limit theorem for dependent non-identically distributed random variables*. Cowles Foundation Discussion Paper, Yale University.
- Baldi, P. (1988). Linear learning: landscapes and algorithms. In Touretzky, D. S. (ed.), *Advances in Neural Information Processing Systems I*, Proceedings of the 1988 NIPS Conference, Denver. Morgan Kaufmann.
- Baldi, P., & Hornik, K. (1991). Back-propagation and unsupervised learning in linear networks. In Chauvin, Y., and Rumelhart, D. E. (eds.), *Back Propagation: Theory, Architectures and Applications*. Earlbaum Associates.
- Barlow, H. B., & Földiák, P. (1989). Adaptation and decorrelation in the cortex. In Miall, C., Durbin, R. M., & Mitchison, G. J. (eds.), *The Computing Neuron*. Addison-Wesley.
- Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, **59**, 291–294.
- Dunford, N., & Schwartz, J. T. (1966). *Linear operators*, Part 1. New York: Wiley.
- Földiák, P. (1989). Adaptive network for optimal linear feature extraction. In *Proceedings of the International Joint Conference on Neural Networks* (pp. I: 401–405). San Diego: SOS Printing.
- Gallant, A. R. & White, H. (1988). *A unified theory of estimation and inference for nonlinear dynamic models*. Oxford: Basil Blackwell.
- Kohonen, T. (1984). *Self-organization and associative memory*. New York: Springer Verlag.
- Krogh, A. & Hertz, J. A. (1990). Hebbian learning of principal components. In Eckmiller, R., Hartmann, G., and Hauske, G. (eds.), *Parallel Processing in Neural Systems and Computers* (pp. 183–186). Elsevier Science Publishers B.V. (North-Holland).
- Kuan, C.-M., & Hornik, K. (1990). *Convergence of learning algorithms with constant learning rates*. Preprint, Department of Economics, University of Illinois at Urbana-Champaign.
- Kushner, H. J., & Clark, D. S. (1978). *Stochastic approximation methods for constrained and unconstrained systems*. New York: Springer Verlag.
- Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, **21**, 105–117.
- Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, **AC-22**, 551–575.
- Ljung, L., & Söderström, T. (1983). *Theory and Practice of Recursive Identification*. Cambridge: MIT Press.

- Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons.
- McLeish, D. L. (1975). A maximal inequality and dependent strong laws. *Annals of Probability*, **3**, 829–839.
- Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematics and Biology*, **15**, 267–273.
- Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, **1**, 61–68.
- Oja, E., & Karhunen, J. (1985). On stochastic approximation of the eigenvectors and the eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, **106**, 69–84.
- Rubner, J. & Tavian, P. (1990). *A self-organizing network for principal component analysis*. Preprint, Physics Department, Technische Universität München, Munich, Germany.
- Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, **2**, 459–473.
- Williams, R. J. (1985). *Feature discovery through error-correction learning*. Technical Report 8501, Institute of Cognitive Science, University of California, San Diego.

KMAN
RY INC.



JUN 95

UNIVERSITY OF ILLINOIS-URBANA



3 0112 060295919